ILLUSTRATIVE EXAMPLES OF PRINCIPAL COMPONENTS ANALYSIS

W. T. Federer, C. E. McCulloch and N. J. Miles-McDermott
Biometrics Unit, Cornell University, Ithaca, New York 14853-7801

BU-901-MA December 1986

Running Title: Principal Component Analysis

CONTACT: Prof. C. E. McCulloch
Biometrics Unit
337 Warren Hall
Cornell University
Ithaca, NY 14853
USA

ABSTRACT

In order to provide a deeper understanding of the workings of principal components, four data sets were constructed by taking linear combinations of values of two uncorrelated variables to form the X-variates for the principal component analysis. These examples are believed useful for aiding researchers in the interpretation of data and they highlight some of the properties and limitations of principal components analyses.

INTRODUCTION

Principal components is a form of multivariate statistical analysis and is one method of studying the correlation or covariance structure in a set of measurements on m variables for n individuals. For example, a data set may consist of n = 260 samples and m = 15 different fatty acid variables. It may be advantageous to study the structure of the 15 fatty acid variables since some or all of the variables may be measuring the same response. One simple method of studying the correlation structure is to compute the m(m-1)/2 pairwise correlations and note which correlations are close to unity. When a group of variables are all highly intercorrelated, one may be selected for use and the others discarded or the sum of all the variables can be used. When the structure is more complex, the method of principal components analysis (PCA) becomes useful.

In order to use and interpret a principal components analysis there needs to be some practical meaning associated with the various principal components. In this paper, we describe the basic features of principal components and examine some constructed examples to illustrate the interpretations that are possible.

BASIC FEATURES OF PRINCIPAL COMPONENTS ANALYSIS

PCA can be performed on either the variances and covariances among the m variable or their correlations. One should always check which is being used in a particular computer package program. First we will consider analyses using the matrix of variances and covariances. A PCA generates m new variables, the principal components (PCs), by forming linear combinations of the original variables, $\mathbf{X} = (X_1, X_2, \cdots, X_m)$, as follows:

$$
\begin{aligned}
PC_1 &= b_{11}X_1 + b_{12}X_2 + \cdots + b_{1m}X_m = \mathbf{X}\mathbf{b}_1 \\
PC_2 &= b_{21}X_1 + b_{22}X_2 + \cdots + b_{2m}X_m = \mathbf{X}\mathbf{b}_2 \\
&\vdots \\
PC_m &= b_{m1}X_1 + b_{m2}X_2 + \cdots + b_{mm}X_m = \mathbf{X}\mathbf{b}_m \ .
\end{aligned}
$$

In matrix notation,

$$\mathbf{P} = (PC_1, PC_2, \cdots, PC_m) = \mathbf{X}(\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_m) = \mathbf{X}\mathbf{B} \ .$$

The rationale in the selection of the coefficients, b_{ij}, that define the linear combinations that are the PC_1 is to try to capture as much of the variation in the original variables with as few PCs as possible. Since the variance of a linear combination of the Xs can be made arbitrarily large by selecting very large coefficients, the b_{ij} are constrained by convention so that the sum of squares of the coefficients for any PC is unity:

$$\Sigma_{j=1}^{m} b_{ij}^2 = 1 \qquad i = 1,2,\cdots,m \ .$$

Under this constraint, the b_{1j} in PC_1 are chosen so that PC_1 has maximal variance among all PCs.

If we denote the variance of X_i by s_i^2 and if we define the total variance, T, as $T = \Sigma_{i=1}^{m} s_i^2$, then the proportion of the variance in the original variables that is captured in PC_1 can be quantified as $var(PC_1)/T$. In selecting the coefficients for PC_2, they are further constrained by the requirement that PC_2 be uncorrelated with PC_1. Subject to this constraint and the constraint that the squared coefficients sum to one, the coefficients b_{2j} are selected so as to maximize $var(PC_2)$. Further coefficients and PCs are selected in a similar manner, by requiring that a PC be uncorrelated with all PCs previously selected and then selecting the coefficients to maximize variance. In this manner, all the PCs are

constructed so that they are uncorrelated and so that the first few PCs capture as much variance as possible. The coefficients also have the following interpretation which helps to relate the PCs back to the original variables. The correlation between the i^{th} PC and the j^{th} variable is

$$b_{ij}\sqrt{var(PC_i)}/s_j \quad .$$

After all m PCs have been constructed, the following identity holds:

$$var(PC_1) + var(PC_2) + \cdots + var(PC_m) = T = \sum_{i=1}^{m} s_i^2 \quad .$$

This equation has the interpretation that the PCs divide up the total variance of the Xs completely. It may happen that one or more of the last few PCs have variance zero. In such a case, all the variation in the data can be captured by fewer than m variables. Actually, a much stronger result is also true; the PCs can also be used to reproduce the actual values of the Xs, not just their variance. We will demonstrate this more explicitly later.

The above properties of PCA are related to a matrix analysis of the variance-covariance matrix of the Xs, $\mathbf{S}_x$. Let $\mathbf{D}$ be a diagonal matrix with entries being the eigenvalues, λ_i, of $\mathbf{S}_x$ arranged in order from largest to smallest. Then the following properties hold:

(i) $\lambda_i = var(PC_i)$

(ii) $trace(\mathbf{S}) = \Sigma_{i=1}^{m} s_i^2 = T = \Sigma_{i=1}^{m} \lambda_i = \Sigma_{i=1}^{m} var(PC_i)$

(iii) $corr(PC_i, X_j) = \dfrac{b_{ij}\sqrt{\lambda_i}}{s_j}$

(iv) $\mathbf{S}_x = \mathbf{B'DB}$,

where $\mathbf{B}$ is the matrix of the coefficients of PCs defined earlier.

The statements made above are for the case when the analysis is performed on the variance-covariance matrix of the Xs. The correlation matrix could also be used, which is equivalent to performing a PCA on the variance-covariance matrix of the standardized variables,

$$Y_i = \frac{X_i - \bar{X}_i}{s_i} \quad .$$

PCA using the correlation matrix is different in these respects:

(i) The total "variance" is m, the number of variables. (It is not truly variance anymore.)

(ii) The correlation between PC_i and X_j is given by $b_{ij}\sqrt{var(PC_i)} = b_{ij}\sqrt{\lambda_i}$. Thus PC_i is most highly correlated with the X_j having the largest coefficient in PC_i in absolute value.

The experimenter must choose whether to use standardized (PCA on a correlation matrix) or unstandardized coefficients (PCA on a variance-covariance matrix). The latter is used when the variables are measured on a comparable basis. This usually means that the variables must be in the same units and have roughly comparable variances. If the variables are measured in different units then the analysis will usually be performed on the standardized scale, otherwise the analysis may only reflect the different scales of measurement. For example, if a number of fatty acid analyses are made, but the variances, s_i^2, and means, $\bar{X}_i$, are obtained on different bases and by different methods, then standardized variables would be used (PCA on the correlation matrix). A note of caution in using standardized variables is that this makes the ranges of all variables comparable and this may not be what an investigator desires.

To illustrate some of the above ideas, a number of examples have been constructed and these are described in the next section. In each case, two variables, Z_1 and Z_2, which are uncorrelated, are used to construct X_i. Thus, all the variance can be captured with two variables and hence only two of the PCs will have nonzero variances. In matrix analysis terms, only two eigenvalues will be nonzero. An important thing to note is that, in general, PCA *will not recover the original variables* Z_1 *and* Z_2. Both standardized and nonstandardized computations will be made.

To compute residuals after computing a particular PC, we proceed as follows. The residuals after computing PC_1 are:

$$\mathbf{X} - \mathbf{X}\mathbf{b}_1\mathbf{b}_1' = \hat{\mathbf{e}}_1 \quad ;$$

those residuals after fitting both the first and second PCs are:

$$\mathbf{X} - \mathbf{X}\mathbf{b}_1\mathbf{b}_1' - \mathbf{X}\mathbf{b}_2\mathbf{b}_2' = \hat{\mathbf{e}}_{12} \quad ;$$

and so forth. The fitting of residuals is also a convenient way of envisioning the extraction of further PCs. PC_2 can be calculated by extracting the *first* PC from the residuals, e_1. That is, it can be calculated by forming the variance-covariance matrix of the residuals and performing PCA on it.

The computations to calculate the residuals are performed on each of the n observations consisting of m variables. A study of residuals may be helpful in finding outliers and patterns of observations.

EXAMPLES

Throughout the examples we will use the variables Z_1 and Z_2 (with n = 11) from which we will construct $X_1, X_2, \ldots, X_m$. We will perform PCA on the Xs. Thus, in our constructed examples, there will only really be two underlying variables.

Values of Z_1 and Z_2

Z_1	-5	-4	-3	-2	-1	0	1	2	3	4	5
Z_2	15	6	-1	-6	-9	-10	-9	-6	-1	6	15

Notice that Z_1 exhibits a linear trend through the 11 samples and Z_2 exhibits a quadratic trend. They are also chosen to have mean zero and be uncorrelated. Z_1 and Z_2 have the following variance-covariance matrix (a variance-covariance matrix has the variance for the i^{th} variable in the i^{th} row and i^{th} column and the covariance between the i^{th} variable and the j^{th} variable in the i^{th} row and j^{th} column).

Variance-covariance matrix of Z_1 and Z_2

$$\begin{pmatrix} 1 & 0 \\ 0 & 85.8 \end{pmatrix}$$

Thus the variance of Z_1 is 11 and the covariance between Z_1 and Z_2 is zero. Also, the total variance is 11 + 85.8 = 96.8.

Example 1: In this first example we analyze Z_1 and Z_2 as if they were the data. Thus $X_1 = Z_1$ and $X_2 = Z_2$ and m = 2. If PCA is performed on the variance-covariance matrix then the output is as follows:

	EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
PRIN1	85.80000[3)]	74.80000	0.88636	0.88636
PRIN2	11.00000[4)]		0.11364	1.00000

EIGENVECTORS

	PRIN1	PRIN2
X1	0.000000[1)]	1.000000[2)]
X2	1.000000	0.000000

The above output was produced using SAS/PRINCOMP, Version 82.3. Note that PRIN1 = PC_1 in the previous notation. We can interpret the results as follows:

1) The first principal component is $PC_1 = 0 \cdot X_1 + 1 \cdot X_2 = X_2$

2) $PC_2 = 1 \cdot X_1 + 0 \cdot X_2 = X_1$

3) $Var(PC_1)$ = eigenvalue = 85.8 = $Var(X_2)$

4) $Var(PC_2)$ = eigenvalue = 11.0 = $Var(X_1)$

The PCs will be the same as the Xs whenever the Xs are uncorrelated. Since X_2 has the larger variance it becomes the first principal component.

If PCA is performed on the correlation matrix we get slightly different results.

Correlation Matrix of Z_1 and Z_2

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

A correlation matrix always has unities along its diagonal and the correlation between the i^{th} variable and the j^{th} variable in the i^{th} row and j^{th} column. PCA would yield the following output:

	EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
PRIN1	1.000000	0.000000	0.500000	0.500000
PRIN2	1.000000		0.500000	1.000000

EIGENVECTORS

	PRIN1	PRIN2
X1	1.000000	0.000000
X2	0.000000	1.000000

The principal components are again the Xs themselves, but both eigenvalues (var(PC)s) are unity since the variables have been first.

Example 2: Let $X_1 = Z_1$, $X_2 = 2Z_1$ and $X_3 = Z_2$. The summary statistics are given below.

	X1	X2	X3
MEAN	0.000000	0.000000	0.000000
ST DEV	3.316625	6.63325	9.262829

CORRELATIONS

	X1	X2	X3
X1	1.0000	1.0000	0.0000
X2	1.0000	1.0000	0.0000
X3	0.0000	0.0000	1.0000

COVARIANCES

	X1	X2	X3
X1	11	22	0
X2	22	44	0
X3	0	0	85.8

TOTAL VARIANCE = 140.8

Note that on the above SAS output, CORRELATIONS means correlation matrix and COVARIANCE means variance-covariance matrix. If the analysis is performed on the variance-covariance matrix the results are shown in Table 1.

Table 1. SAS Output from PCA of Data from Example 2

(Variance-Covariance)

	EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
PRIN1	85.80000	30.80000	0.60938	0.60938
PRIN2	55.00000	55.00000	0.39062	1.00000
PRIN3	0.00000		0.00000	1.00000

EIGENVECTORS

	PRIN1	PRIN2	PRIN3
X1	0.000000	0.447214	0.894427
X2	0.000000	0.894427	-.447214
X3	1.000000	0.000000	0.000000

OBS	X1	X2	X3	PRIN1	PRIN2	PRIN3
1	-5	-10	15	15	-11.180	0
2	-4	-8	6	6	-8.944	0
3	-3	-6	-1	-1	-6.708	0
4	-2	-4	-6	-6	-4.472	0
5	-1	-2	-9	-9	-2.236	0
6	0	0	-10	-10	0.000	0
7	1	2	-9	-9	2.236	0
8	2	4	-6	-6	4.472	0
9	3	6	-1	-1	6.708	0
10	4	8	6	6	8.944	0
11	5	10	15	15	11.180	0

The values under PRIN1 and PRIN2 are calculated by evaluating the PCs for each sample. For example,

$$-11.180 = .447214\ X_1 + .894427\ X_2 + 0 \cdot X_3$$

$$= .447214(-5) + .894427(-10) + 0 \cdot (15)$$

Analyzing the correlation matrix gives the results in Table 2.

Table 2. SAS Output from PCA of Data for Example 2 (Correlation).

	EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
PRIN1	2.000000	1.000000	0.666667	0.666667
PRIN2	1.000000	1.000000	0.333333	1.000000
PRIN3	0.000000		0.000000	1.000000

EIGENVECTORS

	PRIN1	PRIN2	PRIN3
X1	0.707107	0.000000	-.707107
X2	0.707107	0.000000	0.707107
Z	0.000000	1.000000	0.000000

OBS	X1	X2	X3	PRIN1	PRIN2	PRIN3
1	-5	-10	15	-2.1320	1.6194	0
2	-4	-8	6	-1.7056	0.6478	0
3	-3	-6	-1	-1.2792	-0.1080	0
4	-2	-4	-6	-0.8528	-0.6478	0
5	-1	-2	-9	-0.4264	-0.9716	0
6	0	0	-10	0.0000	-1.0796	0
7	1	2	-9	0.4264	-0.9716	0
8	2	4	-6	0.8528	-0.6478	0
9	3	6	-1	1.2792	-0.1080	0
10	4	8	6	1.7056	0.6478	0
11	5	10	15	2.1320	1.6194	0

The values for PRIN1, PRIN2 and PRIN3 are calculated in a fashion similar to the variance-covariance analysis, but using the standardized variables. For example,

$$-2.1320 = .707107\left(\frac{X_1-\bar{X}_1}{s_1}\right) + .707107\left(\frac{X_2-\bar{X}_2}{s_2}\right) + 0\left(\frac{X_3-\bar{X}_3}{s_3}\right)$$

$$= .707107\left(\frac{-5-0}{3.316625}\right) + .70107\left(\frac{-10-0}{6.63325}\right) + 0\left(\frac{15-0}{9.262829}\right).$$

There are several items to note in these analyses:

i) There are only two nonzero eigenvalues since given X_1 and X_3, X_2 is computed from X_1.

ii) X_3 is its own principal component since it is uncorrelated with all the other variables.

iii) The sum of the eigenvalues is the sum of the variances, i.e.,

$$11 + 44 + 85.8 = 55 + 85.8 = 140.8$$

and

$$1 + 1 + 1 = 2 + 1 = 3 .$$

iv) For the variance-covariance analysis, the ratio of the coefficients of X_1 and X_2 in PC_2 is the same as the ratio of the variables themselves (since $X_2 = 2X_1$).

v) Since there are only two nonzero eigenvalues, only two of the PCs have nonzero variances (are nonconstant).

vi) The coefficients help to relate the variables and the PCs. In the variance-covariance analysis,

$$\text{Corr}(PC_2,X_1) = \frac{(\text{coefficient of } X_1 \text{ in } PC_2)\sqrt{\text{var}(PC_2)}}{\sqrt{\text{var}(X_1)}}$$

$$= \frac{b_{21}\sqrt{\lambda_2}}{s_1}$$

$$= \frac{.447214\sqrt{55}}{3.16625}$$

$$= 1 .$$

In the correlation analysis,

$$\text{Corr}(PC_1,X_1) = b_{11}\sqrt{\lambda_1}$$

$$= .707107\sqrt{2}$$

$$= 1 .$$

Thus, in both these cases, the variable is perfectly correlated with the PC.

vii) The Xs can be reconstructed exactly from the PCs with nonzero eigenvalues. For example, in the variance-covariance analysis, X_3 is clearly given by PC_1. X_1 and X_2 can be recovered via the formulas

$$X_1 = PC_2/\sqrt{5}$$

$$X_2 = 2 \cdot PC_2/\sqrt{5} \ .$$

As a numerical example,

$$-5 = -11.180/\sqrt{5} \ .$$

Example 3: For Example 3 we use $X_1 = Z_1$, $X_2 = 2(Z_1+5)$, $X_3 = 3(Z_1+5)$ and $X_4 = Z_2$. Thus X_1, X_2 and X_3 are all created from Z_1. The data and summary statistics are given in Table 3.

Table 3. SAS Output of Data, Means, Standard Deviations (ST DEV) Correlation Matrix, and Variance-Covariance Matrix for Example 3.

OBS	X1	X2	X3	X4
1	-5	0	0	15
2	-4	2	3	6
3	-3	4	6	-1
4	-2	6	9	-6
5	-1	8	12	-9
6	0	10	15	-10
7	1	12	18	-9
8	2	14	21	-6
9	3	16	24	-1
10	4	18	27	6
11	5	20	30	15

	X1	X2	X3	X4
MEAN	0.000000	10.00000	15.00000	0.00000
ST DEV	3.316625	6.63325	9.94987	9.62823

CORRELATION MATRIX

	X1	X2	X3	X4
X1	1.0000	1.0000	1.0000	0.0000
X2	1.0000	1.0000	1.0000	0.0000
X3	1.0000	1.0000	1.0000	0.0000
X4	0.0000	0.0000	0.0000	1.0000

COVARIANCES

	X1	X2	X3	X4
X1	11	22	33	0
X2	22	44	66	0
X3	33	66	99	0
X4	0	0	0	85.8

TOTAL VARIANCE = 239.8 = 11 + 44 + 99 + 85.8

The analyses for the variance-covariance matrix (unstandardized analysis) and correlation matrix (standardized analysis) are given in Table 4.

Table 4. SAS Output from PCA of Data for Example 3.

VARIANCE-COVARIANCE ANALYSIS

	EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
PRIN1	154.0000	68.2000	0.6422	0.6422
PRIN2	85.8000	85.8000	0.3578	1.0000
PRIN3	0.0000	0.0000	0.0000	1.0000
PRIN4	0.0000		0.0000	1.0000

EIGENVECTORS

	PRIN1	PRIN2	PRIN3	PRIN4
X1	0.267261	0.000000	0.358569	0.894427
X2	0.534522	0.000000	0.717137	-.447214
X3	0.801784	0.000000	-.597614	0.000000
X4	0.000000	1.000000	0.000000	0.000000

OBS	X1	X2	X3	X4	PRIN1	PRIN2	PRIN3	PRIN4
1	-5	0	0	15	-1.336	15	-1.793	-4.472
2	-4	2	3	6	2.405	6	-1.793	-4.472
3	-3	4	6	-1	6.147	-1	-1.793	-4.472
4	-2	6	9	-6	9.889	-6	-1.793	-4.472
5	-1	8	12	-9	13.630	-9	-1.793	-4.472
6	0	10	15	-10	17.372	-10	-1.793	-4.472
7	1	12	18	-9	21.114	-9	-1.793	-4.472
8	2	14	21	-6	24.855	-6	-1.793	-4.472
9	3	16	24	-1	28.597	-1	-1.793	-4.472
10	4	18	27	6	32.339	6	-1.793	-4.472
11	5	20	30	15	36.080	15	-1.793	-4.472

CORRELATION ANALYSIS

	EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
PRIN1	3.000000	2.000000	0.750000	0.750000
PRIN2	1.000000	1.000000	0.250000	1.000000
PRIN3	0.000000	0.000000	0.000000	1.000000
PRIN4	-.000000		-.000000	1.000000

EIGENVECTORS

	PRIN1	PRIN2	PRIN3	PRIN4
X1	0.577350	-.000000	0.408248	-.707107
X2	0.577350	0.000000	0.408248	0.707107
Z	-.000000	1.000000	-.000000	0.000000
X3	0.577350	0.000000	-.816497	0.000000

Table 4 (Continued)

OBS	X1	X2	X3	X4	PRIN1	PRIN2	PRIN3	PRIN4
1	-5	0	0	15	-2.6112	1.6194	0	0
2	-4	2	3	6	-2.0889	0.6378	0	0
3	-3	4	6	-1	-1.5667	-0.1080	0	0
4	-2	6	9	-6	-1.0445	-0.6478	0	0
5	-1	8	12	-9	-0.5222	-0.9716	0	0
6	0	10	15	-10	0.0000	-1.0796	0	0
7	1	12	18	-9	0.5222	-0.9716	0	0
8	2	14	21	-6	1.0445	-0.6478	0	0
9	3	16	24	-1	1.5667	-0.1080	0	0
10	4	18	27	6	2.0889	0.6478	0	0
11	5	20	30	15	2.6112	1.6194	0	0

For the variance-covariance analysis, the coefficients in PC_1 are in the same ratio as their relationship to Z_1. In the correlation analysis X_1, X_2 and X_3 have equal coefficients. In both analyses, as expected, the total variance is equal to the sum of the variances for the PCs. In both cases two PCs, PC_3 and PC_4, have zero variance; in the correlation analysis the PCs are identically zero but in the variance-covariance analysis they have a constant value.

Example 4: In this example we take more complicated combinations of Z_1 and Z_2. Let

$$X_1 = Z_1$$
$$X_2 = 2Z_1$$
$$X_3 = 3Z_1$$
$$X_4 = Z_1/2 + Z_2$$
$$X_5 = Z_1/4 + Z_2$$
$$X_6 = Z_1/8 + Z_2$$
$$X_7 = Z_2$$

Note that X_1, X_2 and X_3 are collinear (they all have correlation unity) and X_4, X_5, X_6 and X_7 have steadily decreasing correlations with X_1. The data and data summaries are given in Table 5.

Table 5. SAS Output of Data, Means, and Standard Deviation for Data from Example 4.

OBS	X1	X2	X3	X4	X5	X6	X7
1	-5.000	-10.000	-15.000	12.500	13.750	14.375	15.000
2	-4.000	-8.000	-12.000	4.000	5.000	5.500	6.000
3	-3.000	-6.000	-9.000	-2.500	-1.750	-1.375	-1.000
4	-2.000	-4.000	-6.000	-7.000	-6.500	-6.250	-6.000
5	-1.000	-2.000	-3.000	-9.500	-9.250	-9.125	-9.000
6	0.000	0.000	0.000	-10.000	-10.000	-10.000	-10.000
7	1.000	2.000	3.000	-8.500	-8.755	-8.875	-9.000
8	2.000	4.000	6.000	-5.000	-5.500	-5.750	-6.000
9	3.000	6.000	9.000	0.500	-0.250	-0.625	-1.000
10	4.000	8.000	12.000	8.000	7.000	6.500	6.000
11	5.000	10.000	15.000	17.500	16.250	15.625	15.000

	X1	X2	X3	X4	X5	X6	X7
MEAN	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
ST DEV	3.31662	6.63325	9.94987	9.41010	9.29987	9.27210	9.26283

The PCAs for the variance-covariance and correlation matrices are in Table 6.

Table 6. SAS Output of PCAs for Data from Example 4.

COVARIANCES

	X1	X2	X3	X4	X5	X6	X7
X1	11.00000	22.00000	33.00000	5.50000	2.75000	1.37500	0.00000
X2	22.00000	44.00000	66.00000	11.00000	5.50000	2.75000	0.00000
X3	33.00000	66.00000	99.00000	16.50000	8.25000	4.12500	0.00000
X4	5.50000	11.00000	16.50000	88.55000	87.17500	86.48750	85.80000
X5	2.75000	5.50000	8.25000	87.17500	86.48750	86.14375	85.80000
X6	1.37500	2.75000	4.12500	86.48750	86.14375	85.97188	85.80000
X7	0.00000	0.00000	0.00000	85.80000	85.80000	85.80000	85.80000

	EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
PRIN1	347.01507	193.22077	0.69291	0.69291
PRIN2	153.79430	347.01507	0.30709	1.00000
PRIN3	0.00000	0.00000	0.00000	1.00000
PRIN4	0.00000	0.00000	0.00000	1.00000
PRIN5	0.00000	0.00000	0.00000	1.00000
PRIN6	0.00000	0.00000	0.00000	1.00000
PRIN7	0.00000	0.00000	0.00000	1.00000

EIGENVECTORS

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7
X1	0.02502	0.26479	0.96362	0.01243	0.00000	0.01806	0.01450
X2	0.05004	0.52957	-0.14825	0.02485	0.83205	0.03613	0.02899
X3	0.07505	0.79436	-0.22237	0.03728	-0.55470	0.05419	0.04349
X4	0.50482	0.02744	0.00000	-0.70228	0.00000	-0.29061	-0.40835
X5	0.49856	-0.03876	0.00000	0.71035	0.00000	-0.28719	-0.40354
X6	0.49544	-0.07186	0.00000	-0.00326	0.00000	-0.28623	0.81697
X7	0.49231	-0.10495	0.00000	-0.00481	0.00000	0.86404	-0.00508

OBS	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7
1	25.921	-21.332	0	0	0	0	0
2	8.790	-15.937	0	0	0	0	0
3	-4.359	-10.918	0	0	0	0	0
4	-13.525	-6.275	0	0	0	0	0
5	-18.709	-2.009	0	0	0	0	0
6	-19.911	1.881	0	0	0	0	0
7	-17.131	5.395	0	0	0	0	0
8	-10.368	8.533	0	0	0	0	0
9	0.377	11.294	0	0	0	0	0
10	15.104	13.679	0	0	0	0	0
11	33.813	15.688	0	0	0	0	0

Table 6 (Continued)

CORRELATIONS

	X1	X2	X3	X4	X5	X6	X7
X1	1.00000	1.00000	1.00000	0.17623	0.08916	0.04471	0.00000
X2	1.00000	1.00000	1.00000	0.17623	0.08916	0.04471	0.00000
X3	1.00000	1.00000	1.00000	0.17623	0.08916	0.04471	0.00000
X4	0.17623	0.17623	0.17623	1.00000	0.99614	0.99124	0.98435
X5	0.08916	0.08916	0.08916	0.99614	1.00000	0.99901	0.99602
X6	0.04471	0.04471	0.04471	0.99124	0.99901	1.00000	0.99900
X7	0.00000	0.00000	0.00000	0.98435	0.99602	0.99900	1.00000

	EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
PRIN1	4.05217	1.10433	0.57888	0.57888
PRIN2	2.94783	2.94783	0.42112	1.00000
PRIN3	0.00000	1.00000	0.00000	1.00000
PRIN4	0.00000	1.00000	0.00000	1.00000
PRIN5	0.00000	1.00000	0.00000	1.00000
PRIN6	0.00000	1.00000	0.00000	1.00000
PRIN7	0.00000	0.00000	0.00000	1.00000

EIGENVECTORS

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7
X1	0.14429	0.55733	-0.01047	-0.03822	-0.40825	0.01819	0.70711
X2	0.14429	0.55733	-0.01047	-0.03822	-0.40825	0.01819	-0.70711
X3	0.14429	0.55733	-0.01047	-0.03822	0.81650	0.01819	0.00000
X4	0.49334	-0.06831	-0.00052	0.86715	0.00000	0.00121	0.00000
X5	0.48633	-0.11881	0.70763	-0.28505	0.00000	-0.40912	0.00000
X6	0.48133	-0.14409	-0.70635	-0.28505	0.00000	-0.40912	0.00000
X7	0.47535	-0.16918	0.00134	-0.28490	0.00000	0.81501	0.00000

OBS	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7
1	2.238	-3.284	0	0	0	0	0
2	0.543	-2.304	0	0	0	0	0
3	-0.737	-1.432	0	0	0	0	0
4	-1.600	-0.668	0	0	0	0	0
5	-2.048	-0.011	0	0	0	0	0
6	-2.080	0.538	0	0	0	0	0
7	-1.695	0.980	0	0	0	0	0
8	-0.895	1.314	0	0	0	0	0
9	0.321	1.540	0	0	0	0	0
10	1.953	1.658	0	0	0	0	0
11	4.001	1.669	0	0	0	0	0

We note several things:

i) In both analyses there are only two eigenvalues that are nonzero indicating that only two variables are needed. This is not readily apparent from the correlation or variance-covariance matrix.

ii) In PC_1, PC_2, PC_3, PC_4, and PC_6 where the standardized X_1, X_2, X_3, X_4, and X_6 are the same, they have the same coefficients.

iii) Neither PCA recovers Z_1 and Z_2. The PCs with nonzero variances have elements of both Z_1 and Z_2 in them, i.e., neither PC_1 or PC_2 is perfectly correlated with one of the Zs.

If one computes residuals for Example 4, there will be no residuals after the second PC since the first two PCs explain all the variance. The residuals from PC_1, and a PCA for the residuals, are given in Table 7. The SAS program for their calculations are given in Appendix 1. Several things are noteworthy:

1. The variance-covariance matrix for $\hat{e}_1$ has one nonzero eigenvalue.
2. PC_1 from the analysis is the same as PC_2 from the original analysis.
3. The total variance is the original total variance minus the variance of the original PC_1.

Table 7. Residuals after PC_1 for Data from Example 4 and the SAS Output Using these Residuals as Data Instead of Original Data.

E1	COL1	COL2	COL3	COL4	COL5	COL6	COL7
ROW1	-5.64854	-11.2971	-16.9454	-0.585338	0.826926	1.5328	2.23893
ROW2	-4.21992	- 8.43984	-12.6597	-0.617727	0.617727	1.14515	1.67266
ROW3	-2.89094	- 5.78188	- 8.67287	-0.29958	0.423133	0.784534	1.14589
ROW4	-1.6616	- 3.3232	- 4.98493	-0.172188	0.243144	0.450945	0.658611 = $\hat{e}_1$
ROW5	-0.531891	- 1.06378	- 1.59586	-0.0551207	0.144385	0.210825	0.210825
ROW6	0.498181	0.996361	1.49434	0.0516225	-0.0730223	-0.135146	-0.197468
ROW7	1.42862	2.85723	4.28568	0.148041	-0.209199	-0.387647	-0.566267
ROW8	2.25942	4.51883	6.77814	0.234135	-0.33077	-0.61312	-0.895572
ROW9	2.99058	5.98116	8.97174	0.309905	-0.437738	-0.811563	-1.18538
ROW10	3.62211	7.24421	10.8665	0.37535	-0.530101	-0.982977	-1.4357
ROW11	4.154	8.30799	12.4623	0.430471	-0.607859	-1.12736	-1.64653

SIMPLE STATISTICS

	COL1	COL2	COL3	COL4	COL5	COL6	COL7
MEAN	0.000000	0.000000	0.000000	0.0000000	0.0000000	0.000000	0.888888
ST DEV	3.283719	6.567437	9.851156	0.3402822	0.4806475	0.8911123	1.301577

COVARIANCES

	COL1	COL2	COL3	COL4	COL5	COL6	COL7
COL1	10.783	21.566	32.348	1.1174	-1.578	-2.926	- 4.274
COL2	21.566	43.131	64.697	2.2348	-3.157	-5.852	- 8.548
COL3	32.348	64.697	97.045	3.3522	-4.735	-8.778	-12.82
COL4	1.1174	2.2348	3.3522	0.11579	-0.1636	-0.3032	- 0.4429
COL5	- 1.578	- 3.157	- 4.735	0.23102	0.23102	0.42831	0.6256
COL6	- 2.926	- 5.852	- 8.778	-0.3032	0.42831	0.79408	1.1599
COL7	- 4.274	- 8.548	-12.82	-0.4429	0.6256	1.1599	1.6941

TOTAL VARIANCE = 153.7943

	EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
PRIN1	153.7943	153.7943	1.0000	1.0000
PRIN2	0.0000	0.0000	0.0000	1.0000
PRIN3	0.0000	0.0000	0.0000	1.0000
PRIN4	0.0000	0.0000	0.0000	1.0000
PRIN5	-0.0000	0.0000	-0.0000	1.0000
PRIN6	-0.0000	0.0000	-0.0000	1.0000
PRIN7	-0.0000	0.0000	-0.0000	1.0000

EIGENVECTORS

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7
COL1	0.264786	-0.251156	-0.019310	0.836265	0.064150	0.138641	0.379158
COL2	0.529573	-0.502312	0.013628	-0.392969	-0.048768	0.556175	-0.030259
COL3	0.794359	0.446382	0.095146	-0.059164	0.003596	-0.395920	0.020439
COL4	0.027439	-0.010654	0.013988	0.007718	0.973360	0.042589	-0.222788
COL5	-0.038758	0.6520600	0.258748	0.120481	-0.036856	0.699033	-0.042929
COL6	-0.071856	-0.216434	0.960808	-0.007869	-0.000523	-0.018358	0.895849
COL7	-0.104954	0.114924	-0.009529	-0.357909	0.211436	0.018358	0.895849

= **PC_2 FROM RAW DATA**

SUMMARY AND DISCUSSION

PCA provides a method of extracting structure from the variance-covariance or correlation matrix. If a multivariate data set is actually constructed in a linear fashion from fewer variables, then PCA will discover that structure. PCA constructs linear combinations of the original data, **X**, with maximal variance, $\mathbf{P} = \mathbf{XB}$. This relationship can be inverted, $\mathbf{PB}^{-1} = \mathbf{X}$, to recover the Xs from the PCs (actually only those PCs with nonzero eigenvalues are needed – see Example 2). Though PCA will often help discover structure in a data set, it does have limitations. It will not necessarily recover the exact underlying variables, even if they were uncorrelated (Example 4). Also, by its construction, PCA is limited to searching for linear structure in the Xs.

Although the motivation for a PCA is to explain variance with a linear combination of the m variables, there are other things that can be accomplished with a PCA. For example, residuals may be computed (see Example 4). These can be studied to investigate their distribution, or to find patterns or outliers. Linear combinations of variables can also be eliminated if one or more eigenvalues are near zero. A study of collinearity or near collinearity among the variables may be important.

There does not seem to be any systematic way to use PCA as a variable selection tool. Since PCA works by calculating linear combinations of all the variables, usually a PC will have a contribution from each of the original variables. Thus, even in cases in which we need only consider a few of the PCs, each one of them often includes each of the original variables with a substantial coefficient.

Readers interested in further reading about PCA should consult a book on multivariate statistical methods. Some good references are Harris (1975), Morrison (1976), and Johnson and Wichern (1982). Federer, McCulloch and Miles-McDermott (1986) have a technical report which describes in detail output from statistical packages that perform PCA.

BIBLIOGRAPHY

Federer, W.T., McCulloch, C.E., and Miles-McDermott, N. (1986). Illustrative Examples of Principal Components Analysis Using SAS/PRINCOMP. BU-918-M in the Biometrics Unit Series.

Harris, R.J. (1975). *A Primer of Multivariate Statistics*. Academic Press, NY.

Johnson, R.A. and Wichern, D.W. (1982). *Applied Multivariate Analysis*. Prentice Hall, Englewood Cliffs,NJ.

Morrison, D.F. (1976). *Multivariate Statistical Methods* (2nd ed.). McGraw-Hill, NY.

APPENDIX 1

```
DATA ONE;
INPUT Z1 Z2;
X1=Z1;
X2=2*Z1;
X3=3*Z1;
X4=(Z1/2)+Z2;
X5=(Z1/4)+Z2;
X6=(Z1/8)+Z2;
X7=Z2;
DROP Z1 Z2;
CARDS;
-5 15
-4 6
-3 -1
-2 -6
-1 -9
 0 -10
 1 -9
 2 -6
 3 -1
 4 6
 5 15
PROC MATRIX PRINT;
V1=.02502/.05004/.07505/.50482/.49856/.49544/.49231;
FETCH X DATA=ONE;
E1=X-((V1*V1')*X')';
OUTPUT E1 OUT=TWO;
PROC PRINCOMP COV DATA=TWO;
PROC MATRIX PRINT;
V1=.02502/.05004/.07505/.50482/.49856/.49544/.49231;
V2=.26479/.529573/.794359/.027439/-.038757/-.071855/-.104954;
FETCH X DATA=ONE;
E2=X-(((V1*V1')*V1')*X')+((V2*V2')*X'))';
OUTPUT E2 OUT=FOUR;
PROC PRINCOMP COV DATA=FOUR;
```